Multi-level Encoder-Decoder Architectures for Image Restoration

Indra Deep Mastan and Shanmuganathan Raman
Indian Institute of Technology Gandhinagar
Gandhinagar, Gujarat, India
{indra.mastan, shanmuga}@iitgn.ac.in

Abstract

Many real-world solutions for image restoration are learning-free and based on handcrafted image priors such as self-similarity. Recently, deep-learning methods that use training data, have achieved state-of-the-art results in various image restoration tasks (e.g., super-resolution and inpainting). Ulyanov et al. bridge the gap between these two families of methods in [29]. They have shown that learning-free methods perform close to the state-of-the-art learning-based methods ($\approx$ 1 PSNR). Their approach benefits from the encoder-decoder network (ed).

In this paper, we propose a framework based on the multi-level extensions of the encoder-decoder network (med) to investigate interesting aspects of the relationship between image restoration and network construction independent of learning. Our framework allows various network structures by modifying the following network components: skip links, cascading of the network input into intermediate layers, a composition of the encoder-decoder subnetworks, and network depth. These handcrafted network structures illustrate how the construction of untrained networks influence the following image restoration tasks: denoising, super-resolution, and inpainting. We also demonstrate image reconstruction using flash and no-flash image pairs. We provide performance comparisons with the state-of-the-art methods for all the restoration tasks above.

1. Introduction

Image restoration is an *ill-posed* problem which aims to recover an image given its corrupted observation (e.g., denoising [39, 5, 31], super-resolution [14, 23, 2], and inpainting [36, 34, 33]). Corruption may occur due to noise, camera shake, and due to the fact that the picture was taken in rain or underwater [16]. Image restoration methods could be mainly classified into two types - *traditional* methods and *deep-learning* (DL) methods. Traditional methods include spatial filtering methods (e.g., bilateral filters [28], non-local means [4]), wavelet transform based methods [6], and dictionary learning and sparse coding [17, 37]. DL methods generally include a neural network to learn image prior from the training samples (learning-based[1]) for restoration, where the training samples contain paired examples of corrupted and high-quality images.

Traditional methods are generally faster and comparatively less cumbersome to implement, *e.g.,* filtering approaches [8]. Whereas DL methods could be tricky to implement. For example, methods based on adversarial loss require training of two separate networks, namely a generator and a discriminator [14]. Moreover, DL methods output photo-realistic images with finer details of features due to the image prior being captured by feature learning on a collection of images [14, 38, 3].

Representation learning from images gives insight into the image statistics captured by the network. The main idea is to perform various image restoration tasks to learn a better image prior [12]. However, it is focused on the learning-based setting [1]. There are fewer studies that directly investigate the image prior captured by the neural network without using any training datasets. Ulyanov *et al.* first conducted the studies to achieve image restoration without using a training sample (learning-free) [29]. This paper focuses on the research thread mentioned above. Our work combine the ideas of traditional methods and the DL approaches similar to [10, 25, 15, 35, 29].

Our ablation study shows how the structure of the untrained network influences the quality of image restoration achieved by them. For example, inpainting of a large missing region is qualitatively better-achieved using an encoder-decoder network *without* skip connections, whereas super-resolution is better-achieved *with* skip links (Fig. 10 and Fig. 11).

We have performed extensive experiments on various handcrafted network architectures obtained by modifying the network components. We focus on the following network components: *depth* of the network, *skip connections*, cascading of the network input into intermediate layers

[1]The learning refers to training the network on the collection of images and learning-free refers to the methods which do not use training data.

(*cascade*), and composition of the encoder-decoder subnetworks (*composition*). We show how each of the above network components affects image restoration. For example, we show how the performance of denoising gets affected when we increase the depth of the network (Fig. 4).

We have formulated a framework called multi-level encoder decoder (med) that models various handcrafted network architectures. An instance from our framework med is a composition of three encoder-decoder networks (Fig. 2). The multi-level extension of encoder-decoder is motivated to exploit the re-occurrence of an image patch at different resolutions. We show our analysis using six different network instances of med (Table 1). These handcrafted network architectures help us develop insight into how the network construction influences image restoration (Fig. 4, Fig. 7, Fig. 8, Fig. 10, and Fig. 11). The key idea is to iteratively minimize the loss between the network output and the corrupted image to implicitly capture the image prior in the network.

There is an inherent contrast in our objectives. On the one hand, we aim to experiment with various high capacity networks to show the relation between image restoration and network construction. The higher depth allows more network components and various network structures for the analysis of the image prior. On the other hand, the high capacity network should not negatively influence the quality of image restoration. This is due to the fact that the higher depth network suffers from the *vanishing gradients* problem [18, 26]. One option is to use skip links to propagate the gradients and feed the image features from the intermediate layers to the last layers of the network [18]. Our main contributions are summarized as follows.

- To the best of our knowledge, this is the first study of a multi-level encoder-decoder framework (med) designed to illustrate the relationship between image restoration and network construction, independent of training data and using DL. The med framework allows analysis of the deep prior by using four networks components (*depth*, *skip connections*, *composition*, *cascade*) whereas DIP [29] includes the investigation based on the two network components (*depth* and *skip connections*). The med framework provides a more rigorous evaluation of the usefulness of skip connections compared to [29].

- We also perform various image restoration tasks to show the quality of the image prior captured by the multi-level network architectures. We have achieved results comparable to the state-of-the-art methods for denoising, super-resolution, and inpainting with $x\%$ pixels drop despite experimenting with various high-capacity networks. We also observe a better flash no-flash based image construction when compared to [29].

2. Related work

Image restoration aims to recover a good quality image from a corrupted observation. It is a useful preprocessing step for other problems, e.g., classification [30]. Mao *et al.* have shown image restoration using an ed network with symmetric skip links between the layers of encoder and decoder [18]. There are various proposals for the loss functions for the image restoration tasks, e.g., adversarial loss [14], perceptual loss [14], or contextual loss [20, 19]. In addition, Chang *et al.* have proposed a single generic network for various image restoration tasks [22]. However, the drawback to this line of work is that the restoration output could be biased toward a particular training dataset.

Ulyanov *et al.* showed that a randomly-initialized ed network works as a hand-crafted prior for restoring images without training data [29]. Motivated by their approach, our learning-free framework only uses the handcrafted structure of the network for image restoration. However, unlike [29], we explore how the network components directly influence various image restoration tasks.

3. Multi-level Encoder-Decoder Framework

In this section, we explain the multi-level encoder-decoder framework (med) and its major components. We shall also discuss an example construction of a multi-level encoder-decoder network and then provide a classification of the networks useful for our experiments.

The med is one of the general class of networks, where each network is a composition of encoder-decoder blocks as subnetworks. We address med as a network $\mathcal{F}$ for devising a simpler explanation. The med network $\mathcal{F}$ is a composition of two subnetworks, namely a *generator* G and an *enhancer* E. The image restoration network $\mathcal{F}$ is defined in Eq. 1.

$$\mathcal{F} = E \circ G \tag{1}$$

Here, the *generator* and the *enhancer* are either an encoder-decoder network (ed) or a composition of ed networks. The encoders determine the abstract representation of the image features, which are used by the decoder for the reconstruction of the image. The composition of networks allows multiple sub-networks to learn image features from the down-sampled versions of the corrupted image. This would enforce the output of the *generator* to be consistent across the multiple scales of the *target* image[2] to improvise the quality of the image restoration.

The multi-level encoder-decoder framework is motivated to model various network architectures by modifying the network components described in Subsection 3.1. For example, let's suppose the *generator* is a depth-k ed network.

[2] *Target* image refers to the high-quality image whose corrupted observation $\hat{I}$ is given for restoration.

There are five network configurations obtained by modifying the skip connections, namely, *Intra-skip*, *Inter-skip encoder-encoder*, *Inter-skip decoder-encoder*, *No-skip*, and *Full-skip* connections[3]. There are two network configurations based on the cascading of the network input, *i.e.*, network with *cascade* or network without *cascade*. There could be $(k-1)$ different *generator-enhancer* compositions for a depth-k *generator* network. We do not consider depth-k *enhancer* to reduce the model capacity. Finally, given a depth-k ed network as the *generator*, the med framework will allow $1\times5\times2\times(k-1)=10(k-1)$ different network structures. On the other hand, [29] will allow only two different network configurations (network with skip connections and without skip connections). Therefore, the generalization med provides various networks to analyze the effects of network components on the quality of the image restoration. Technically, the med is a general framework to explore the nature of the mapping between the network parameter space and the natural image space.

3.1. Network Components

We focus on the following components to show how the network structure affects the image restoration output. (a) *skip* connections, (b) *depth* of the network, (c) *cascading* of the *network input* into the intermediate layers, and (d) *composition* of two ed networks. We describe each of these components as follows.

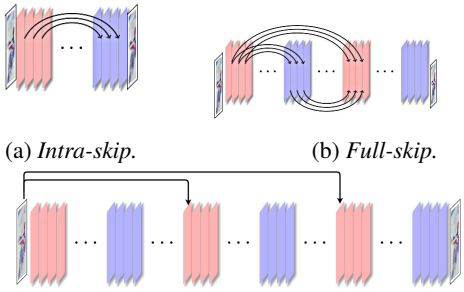
(a) Skip connections. The skip link between the layers L_i and L_j, where i and j are the indices of the network layers with $i<j$, is made by concatenating the output of the layer L_{j-1} with the output of the layer L_i and then feeding into the layer L_j. We have provided the detailed classification of the skip connections in supplementary material. In Fig. 1(a) and Fig. 1(b), we have pictorially shown useful skip link configurations for the paper.

(b) Depth of the network. It is measured by the number of layers present in the network. Higher depth networks capture finer feature details. However, a very high depth could negatively influence the performance (Fig. 3). There are two ways to increase network depth. First, by introducing a new layer into the encoder-decoder (ed) network. Second, by performing a composition of the two ed networks.

(c) Cascading of network input (*cascade*). It is a procedure to successively down-sample the network input and then feed it into the intermediate layers of the network. Formally, to provide the network input at the intermediate layer L, we resize the network input and then concatenate it with the layer $L-1$. Next, we feed the resulting tensor into the layer L. Cascading of network inputs was also utilized by Chen *et al.* [7]. We use it to provide the image features into the *enhancer* network (Fig. 1(c)).

[3]In the supplementary material we have provided the details of different types of skip connections.

(d) Composition of ed networks (*composition*). The composition of two encoder-decoder networks is achieved by feeding the output of the first ed network into the second ed network. The composition of two ed networks increases the network depth and the number of skip connections. The main objective of performing the network composition is to learn image features from the downsampled versions of the corrupted image.



(a) *Intra-skip.* (b) *Full-skip.*

(c) *Cascading* of *network input.*

Figure 1: **Network components.** Layers of the encoder are in *red* and layers of the decoder are in *blue*. (a) *Intra-skip*: the skip connections within EDS network. (b) *Full-skip*: both the *Intra-skip* connections and *Inter-skip* connections are present. (c) Cascading of the network input.

3.2. Multi-level Encoder-Decoder Network

Here, we give an example construction of med network $\mathcal{F}$. It is a three-level ed network where the *generator* is the first ed and the *enhancer* is a composition of the other two ed (Fig. 2).

$$\mathcal{F}=E^2\circ E^1\circ G \tag{2}$$

In Eq. 2, the subnetwork G is the *generator* and subnetwork $E^2\circ E^1$ is the *enhancer* E. The networks G, E^1, and E^2 are defined as follows. $G:\mathbb{R}^{m\times n\times c}\rightarrow\mathbb{R}^{m\times n\times c}$, $E^1:\mathbb{R}^{\frac{m}{2}\times\frac{n}{2}\times c}\rightarrow\mathbb{R}^{\frac{m}{2}\times\frac{n}{2}\times c}$, and $E^2:\mathbb{R}^{\frac{m}{4}\times\frac{n}{4}\times c}\rightarrow\mathbb{R}^{\frac{m}{4}\times\frac{n}{4}\times c}$. Here, c is the number of channels (c is 3 for RGB images). The *generator* G operates at $2\times$ the resolution of E^1 and $4\times$ the resolution of E^2. A resize operator R is used to down-sample the output of G to feed into E^1 and down-sample the output of E^1 to feed into E^2. We have abstracted out R in Eq. 2 for devising a simpler explanation. As described earlier, the *enhancer* $E=E^2\circ E^1$ is mainly used to improvise the output of the *generator* G by making it consistent across different resolutions of the target images.

3.3. Network Classification

We have provided an example construction of a multi-level encoder-decoder network in Fig. 2. Similarly, there are various other network architectures we can get by modifying the network components. We give a classification of

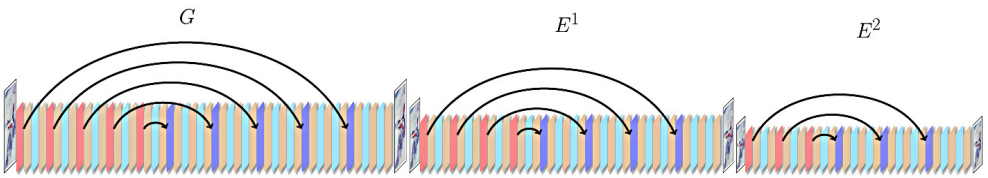


Figure 2: **Multi-level encoder-decoder network architecture.** An example construction of a three-level med network. The *generator* G is an ed network and *enhancer* $E = E^1 \circ E^2$ is the composition of two ed networks. There are skip connections within each ed subnetwork. The layers are shown using colors as follows: `Convolutional layer with stride =1`, `Convolutional layer with stride=2`, `Batch Normalization`, `and Upsampling`. The subnetwork G is a depth-5 ed network, E^1 is a depth-4 ed network, and E^2 is a depth-3 ed network.

med networks useful for *our methods* to analyze these network architectures. The med network is classified based on skip links and cascading of the network input, as shown in Table 1. The network MED has no skip connections and MEDS has *Intra-skip* connections (the character "S" in MEDS denotes the presence of skip connections). The network MEDSF has *Full-skip* connections. Similarly, MEDC has cascading of network input without skip connections (the character "C" in MEDC denotes the cascading of network input). MEDSFC has cascading of network input with *Full-skip* connections. We will use the networks given in Table 1 for our experiments. For example, to see the effects of the decreasing skip links, one could perform image restoration with MEDSF, MEDS, and MED networks.

	No skip	*Intra-skip*	*Full-Skip*
Cascade	MEDC	MEDSC	MEDSFC
No Cascade	MED	MEDS	MEDSF

Table 1: **Classification of med networks.** The classification is based on the following network components: skip-links and cascading of network input at intermediate layers. The graphical representations of the above network components are shown in Fig. 1.

4. Applications

In this section, we show the performance on the following image restoration tasks: super-resolution, denoising, inpainting, and flash no-flash. We provide the technical details of the experiments in the supplementary material.

The aim of image restoration is to reconstruct the image features given a corrupted image $\hat{I}$. The image $\hat{I}$ is computed by adding noise or blur or downsampling the target image I. Ulyanov *et al.* formulated the image restoration problem to the setting of DL based learning-free framework [29]. The image restoration framework is as follows.

$$\theta^* = \underset{\theta}{arg\,min}\ \mathcal{L}(\mathcal{F}_\theta(\hat{z}), \hat{I}); \tag{3}$$

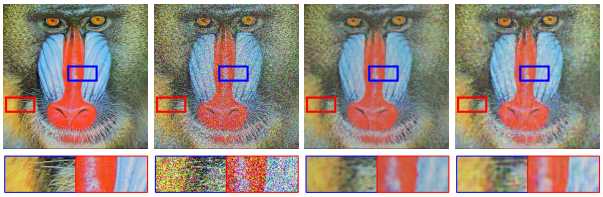
Here, $\mathcal{L}$ is the loss function and $\mathcal{F}$ is a network with parameters denoted by θ and the network input z is prepared from the corrupted image $\hat{I}$. The loss function in the Eq. 3 is a general definition. We now discuss how to perform various image restoration tasks.

Denoising. Denoising aims to reduce noise and recover the clean image where the learning process is assisted only by the corrupted image. Consider a noisy image $\hat{I}$. Let $d_1 = \mathcal{D}(\frac{1}{2}, \hat{I})$ and $d_2 = \mathcal{D}(\frac{1}{4}, \hat{I})$ be the down-sampled versions of the image $\hat{I}$. Our approaches are based on the following property of a natural image: patch recurrence within and across multiple scales. Using this property, one could say that the down-sampled corrupted image contains some of the image features. To make the best use of the property above, our multi-scale loss $\mathcal{L}(\mathcal{F}_\theta(z), \hat{I})$ (Eq. 3) for denoising is defined in Eq. 4.

$$\begin{aligned}\theta^* = \underset{\theta}{arg\,min}\ & \lambda_1 \|G_\theta(z) - \hat{I}\| \\ & + \lambda_2 \|E^1_\theta(z) - d_1\| + \lambda_3 \|E^2_\theta(z) - d_2\|\end{aligned} \tag{4}$$

Here, $\mathcal{F} = E^2 \circ E^1 \circ G$ (Eq. 2). Loss function in Eq. 4 enforces the output of the *generator* to be consistent across the multiple resolutions of the target image. Stated differently, the network performs image restoration at multiple resolutions. Intuitively, achieving restoration at multiple scales is more challenging than at a single scale. Therefore, we expect that solving a harder problem could help in learning a better image prior [12]. The image prior is implicitly captured by the network which is required to restore the image features [29].

Denoising using our MEDSF is shown in Fig. 3. Our MEDSF achieves SSIM=0.72 whereas the baseline DIP [29] outputs a SSIM of 0.71 for a noise strength of $\sigma = 100$. The PSNR values for our MEDSF is 20.95 and DIP outputs a PSNR of 21.36. In Fig. 5, we can observe that a higher PSNR value do not imply higher perceptual quality. We emphasize that the learning-free methods are sensitive



(a) Original, image (b) Noisy, image (c) DIP, (0.479, 18.65) (d) Ours, (0.496, 18.39)

Figure 3: **Denoising**. A comparison between DIP [29] and our MEDSF for denoising with noise strength of $\sigma = 100$ using the performance metric (SSIM, PSNR).

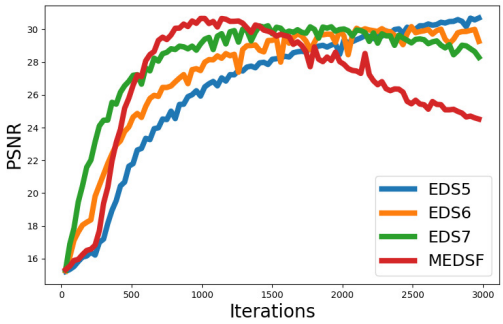


Figure 4: **Network depth effects on denoising.** EDS5 is a depth-5 ed network with skip connections (similarly for EDS6 and EDS7). The highest-depth network MEDSF converges faster. EDS5 network (lower depth) achieves the highest PSNR value but converges the slowest. This shows that a higher model capacity does not necessarily lead to improved performance.

to hyper-parameters[4]. Therefore, the performance of DIP and our MEDSF could probably be further maximized by changing the hyper-parameters.

In Fig. 4, we can observe the effects of network depth on denoising. The network initially learns the global features from the corrupted image by minimizing the loss function defined in Eq. 4. Later, the network starts learning fine feature details which includes noise. Therefore, due to over learning, it produces noisy spots similar to the ones contained in the corrupted image. For example, MEDSF intermediate output at around 1000 iterations is the desired noise free image because it achieves the maximum PSNR.

Super-resolution. Given a *low-resolution* (LR) image $\hat{I} \in \mathbb{R}^{m\times n\times 3}$, and a scaling factor t, super-resolution aims to enhance the image quality and generate a *high-resolution* (HR) image $I^H \in \mathbb{R}^{mt\times nt\times 3}$. We feed network input z into med network $\mathcal{F} = E^2 \circ E^1 \circ G$ and solve the following minimization problem given in Eq. 5.

$$\theta^* = arg\min_{\theta} \lambda_1 \|G_\theta(z) - u_0\| + \lambda_2 \|E^1_\theta(z) - u_1\| + \lambda_3 \|E^2_\theta(z) - \hat{I}\| \tag{5}$$

Here, $u_0 = \mathcal{U}(\hat{I}, 4)$ and $u_1 = \mathcal{U}(\hat{I}, 2)$ are the up-sampled versions of the corrupted LR image $\hat{I}$. Eq. 5 determines the network parameter θ^* which minimizes the loss $\mathcal{L}(\mathcal{F}_\theta(z), \hat{I})$.

Super-resolution achieved by Ulyanov *et al.* in Deep Image Prior (DIP) is the state-of-the-art in DL-based learning-free methods to the best of our knowledge [29]. DIP does not use training samples to learn the image prior in contrast to the learning-based methods which benefit from the training data and *adversarial* loss or *perceptual* loss [23, 14]. Thus, it lacks local level features in the output image. However, it is shown to output better images than various learning-free methods such as bicubic upsampling [29].

We achieved an average SSIM of 0.80, whereas DIP [29] achieved an average SSIM of 0.81 for $4\times$super-resolution. We obtained 24.48 as the average PSNR. Whereas DIP achieved an average PSNR of 25.14[5]. The perceptual quality of the generated images by the proposed approach is observed to be comparable to that of DIP (Fig. 5).

Image inpainting. It involves computing missing pixel values in the corrupted image $\hat{I}$ using the corresponding binary mask $m \in \{0, 1\}^{k\times l}$. Inpainting has various applications such as removing undesirable objects and text in an image, restoring damaged paintings, and computing missing pixels lost during transmission.

Suppose $\mathcal{I}$ is the *target* image and the corrupted image $\hat{I}$ is obtained using the mask m as follows $\hat{I} = \mathcal{I} \odot m$, where $\odot$ is the Hadamard product. Let $d_1 = \mathcal{D}(\frac{1}{2}, \hat{I})$ and $d_2 = \mathcal{D}(\frac{1}{4}, \hat{I})$ be the down-sampled versions of the corrupted image $\hat{I}$, and $m_1 = \mathcal{D}(\frac{1}{2}, m)$ and $m_2 = \mathcal{D}(\frac{1}{4}, m)$ be the down-sampled versions of the mask m. We solve the following minimization problem given in Eq. 6.

$$\theta^* = arg\min_{\theta} \lambda_1 \|(G_\theta(z) - \hat{I}) \odot m\| + \lambda_2 \|(E^1_\theta(z) - d_1) \odot m_2\| + \lambda_3 \|(E^2_\theta(z) - d_2) \odot m_3\| \tag{6}$$

We show the following three inpainting tasks. (1) *restoring missing pixels* lost by masking the target image with a randomly generated binary mask (Fig. 6), (2) *region-inpainting* which includes painting a large region (Fig. 7 and Fig. 10), and (3) *removing text* superimposed on an image (Fig. 8).

[4]The learning-free methods are sensitive to hyper-parameters shown in Fig. 4 of the supplementary material and DIP [29].

[5]RGB images in Set14 dataset had three channels and our med network also outputs RGB images having three channels. However, super-resolution output of DIP [29] have images with four channels (including the *alpha* channel). Therefore, to get a *fair comparison*, we reproduced the DIP output before drawing the comparison.

(a) High resolution image. (b) Low resolution image. (c) DIP [29], (0.88, 28.2). (d) MEDSF, (0.88, 25.43).

Figure 5: $4\times$ **Image super-resolution.** A qualitative comparision using performance metric (SSIM, PSNR). We can observe that a higher PSNR value does not imply a higher perceptual quality.

Inpainting requires understanding the global context and the local structure of the target image [34]. We believe that region-inpainting is the most challenging task because the information from the nearby pixels might not always be sufficient to complete the scene.

We obtained 24.62 as the average PSNR and 0.86 as the average SSIM for inpainting with 90% missing pixels. Whereas DIP [29] achieved an average PSNR of 25.05 and an average SSIM of 0.86. The perceptual quality of the generated images by the proposed approach is observed to be comparable to that of the other methods (Fig. 6).

Flash No-flash. Given a pair of flash and no-flash images, the objective is to get a single high-quality image which incorporates details of the scene from the flash image and ambient illumination from the no-flash image [21, 9]. The combined image helps to achieve denoising, white balancing, red-eye correction [21], foreground extraction [27], and saliency detection [11].

Consider a pair (I^F, I^{NF}), where I^F is a flash image and I^{NF} is a no-flash image. The network input z is prepared by concatenating I^F and I^{NF}. Let $f_1 = \mathcal{D}(\frac{1}{2}, I^{NF})$ and $f_2 = \mathcal{D}(\frac{1}{4}, I^{NF})$ be the down-sampled versions of I^{NF}. We solve the optimization problem given in Eq. 7.

$$\theta^* = \underset{\theta}{arg\,\min}\ \lambda_1 \Big(\|G_\theta(z) - I^{NF}\| + \|E^1_\theta(z) - f_1\| + \|E^2_\theta(z) - f_2\| \Big) + \lambda_2 \|G_\theta(z) - I^F\| \tag{7}$$

Here, λ_1 and λ_2 are the coefficients to control the image features from I^{NF} and I^F. The flash no-flash output is shown in Fig. 9. It is worth noting that our implementation of flash no-flash is more flexible in providing features from both flash and no-flash images using coefficients λ_1 and λ_2, unlike [29] (Fig. 12 of the supplementary material).

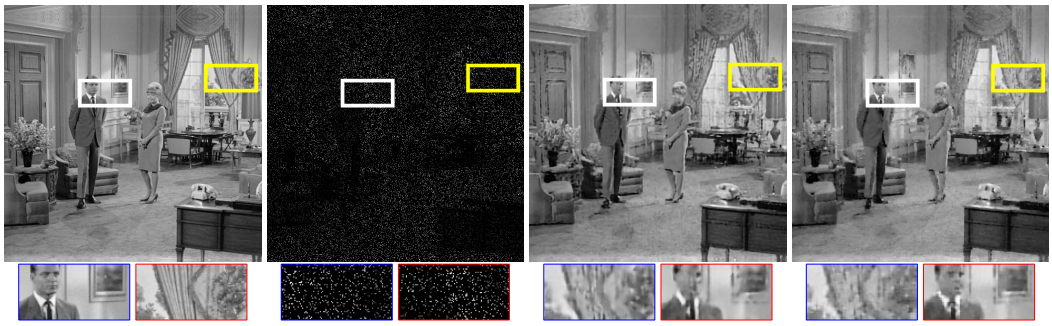
5. Network Structures Effects on Restoration

Here, we discuss the various aspects of the relation between the network construction and image restoration using the med framework. Our choice of the multi-level architecture (a high capacity network) is motivated to illustrate the behavior of various network components (Sec. 3). We emphasize that the image restoration quality from untrained networks is sensitive to hyper-parameters search [29]. We now discuss the results of the ablation studies that we have conducted.

Effects of Skip links. Skip connections have shown *adverse* effects on inpainting, see Fig. 7, Fig. 8, and Fig. 10 (the number of skip connections in the above figures decreases from left to right). Our interpretation of the adverse effects is as follows. The layers of *encoder* have under-developed regions and their pixel values are close to that of the mask (either zero or one). The skip connections pass such intermediate representation to the decoder, which leads to reconstruction bias. Therefore, output images have pixel values that are close to the mask.

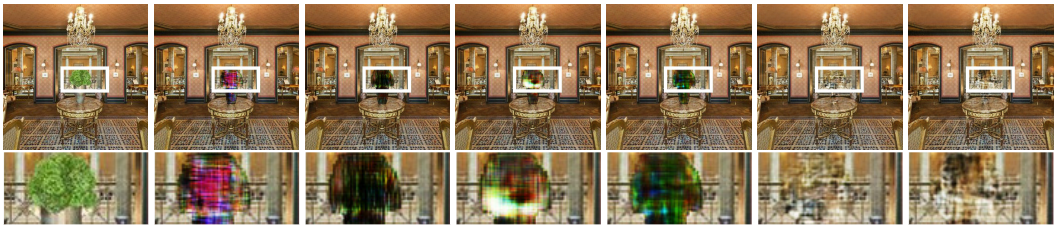
Effects of Depth. In Fig. 4, we observe a higher the depth network converges faster because it has a large number of parameters. However, a lower depth network EDS5 could achieve better restoration than the higher depth network MEDSF. There could be two major factors for the above result. First, higher depth network suffer from the *vanishing gradient* problem which negatively influences the performance [18, 26]. Second, the increase in the number skip connections due to higher depth, influence the performance positively [18]. We believe that the decrease in the PSNR value indicates that the negative influence of the network depth could have more impact compared to the performance enhancement we get from skip connections.

Effects of Cascading of network input (*cascade*). The *cascade* and the *skip connection* looks similar because they both provide image features to the intermediate layers of the network. However, they provide a different type of image features. *Cascade* provides image features from the corrupted image. Whereas, *skip connections* pass the



(a) Original image (b) Corrupted image (c) DIP [29], (0.85, 25.48) (d) Ours, (0.85, 26.18)

Figure 6: **Inpainting.** A comparision for restoration of 90% missing pixels using performance metric (SSIM, PSNR).



(a) Original image (b) MEDSFC, 24.87 (c) MEDSF, 23.20 (d) MEDSC, 24.28 (e) MEDS, 24.04 (f) MEDC, 26.80 (g) MED, 25.83

Figure 7: **Cascading of network input.** Effects of *cascading* of the *network input* in the intermediate layers of the network on removing an object from an image given in (a). The *vase* present in (a) is removed using a white mask, and then inpainting is performed. Networks in (b) and (c) have the same set of skip links. Similarly, (d) and (e) have the same collection of skip links, and (f) and (g) do not have skip links. *Cascading* of *network input* is performed in (b), (d) and (f). (b) and (c) shows that cascading of network input into the intermediate layers of the network improves the performance. Similarly, we can observe that the cascading of network input performed better for other networks: (d) and (e), and (f) and (g).

	DIP [29]	*med* (Ours)
Depth	✔	✔
Skip-links	✔	✔
Composition of *ed*	✖	✔
Cascading of input	✖	✔

Table 2: Network components to investigate the influence of the network structures for image restoration tasks.

	PSNR		SSIM	
	DIP [29]	Ours	DIP [29]	Ours
Denoising	21.36	20.95	0.71	0.72
Inpainting	25.05	24.62	0.86	0.86
SISR	25.14	24.48	0.81	0.80

Table 3: A quantitative comparison for denoising, inpainting, and single image super-resolution (SISR) using average PSNR and SSIM. We provide the visual comparison of generated images in the supplementary material. The perceptual quality of the generated images is comparable to DIP [29] despite the *med* network has a higher capacity to accommodate various network components (Table 2).

image features from the intermediate layers of the network. Object removal (inpainting) is better achieved using *cascade* (Fig. 7). Whereas providing image features using skip connections have shown *adverse* effects for inpainting (Fig. 10). This could be because of the image features captured at the intermediate layers of the network are less interpretable than the features present in the corrupted image.

Effects of Composition of *ed* networks. The two-level *med* network performed better than a three level *med* network for text-removal from an image (Fig. 8). However, the performance difference is not very significant (less than one PSNR). A network composition increases the network depth and the number of skip connections. Therefore, a three-level *med* could have more influence on restoration from skip connections compared to a two-level *med* network. Similarly, a three-level *med* could also increase the effects of *vanishing gradients* due to the higher depth. The composition of the networks shows the combined effects of increasing depth and skip connections.

(a) Masked image (b) MEDSF*, 30.92 (c) MEDSF, 30.40 (d) MEDS*, 31.05 (e) MEDS, 30.35 (f) MED*, 30.18 (g) MED, 29.35

Figure 8: **Composition of networks.** Effects of the composition of the ed networks. MEDSF*, MEDS*, and MED* are two levels ed networks. MED, MEDS, and MEDSF are three level ed networks (*enhancer* is a composition of two ed networks). (b) and (c) shows that the two-level full-skip network performed better than three levels of the full-skip network. Similarly, we can observe that the two-level med network performed better for other networks: (d) and (e), and (f) and (g).

(a) Flash image (b) No-flash (c) DIP, 17.03 (d) Ours, 18.54

Figure 9: **Flash-no flash reconstruction.** (a) Flash image. (b) No flash image. (c) DIP. (d) Ours MEDS.

(a) Masked image. (b) MEDSF. (c) MED.

Figure 10: **Skip connections (I).** The network with skip links (MEDSF) does not perform well for *region inpainting* compared to the network without skip connections.

(a) LR image. (b) MED, 0.63. (c) MEDSF, 0.70.

Figure 11: **Skip connections (II).** Skip links (MEDSF) improves 4× super-resolution as shown by SSIM.

6. Conclusion

We have shown interesting aspects of the relationship between image restoration and network construction. Our methods are unsupervised and they only use the corrupted image for restoration instead of using any training data. Therefore, we believe that it does not produce a biased output unlike learning-based methods, e.g., model collapse [24]. We feel it is a challenging experimental setup compared to supervised learning setup because the network is not learning image features by the pairs of low and high-quality images. The challenge is the limited contextual understanding due to the lack of feature learning from the training data.

Our med framework is a generalization of DIP [29]. This generalization is novel because it incorporates various network components and an *enhancer* network. The med framework is more expressive in terms of casting different network structures to perform the ablation studies for various aspects of the network (Table 2). We also discuss image restoration task specific network structures that perform comparably to the state-of-the-art methods (Table 3).

The major components of the restoration framework are the network and the loss function (Eq. 3). We have shown analysis using various network structures and MSE loss[6]. The study of MSE loss is useful as it is used in other image restoration methods. For example, MSE with adversarial loss in [13, 25] and MSE with contextual loss in [19].

We observed that some network components do not enhance the restoration quality. For example, a network with skip links does not perform well for inpainting. Therefore, the experiments on a network with skip connections for inpainting will not be efficient. Wang *et al.* have used skip connections for video inpainting [32]. However, their approach is in the supervised learning setup, unlike our unsupervised setup. We believe that there are similarities in both setups. For example, if a network component is negatively influencing the image prior learning from the corrupted image (unsupervised setup), then it should also negatively influence the learning from the multiple images of training data (supervised setup). We propose as future work to study our restoration framework in the supervised learning setup.

[6]In Fig. 13 of the supplementary material, we show that MSE performed better than *contextual loss* [20].

References

[1] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. *CVPR*, 2017.

[2] A. Bhowmik, S. Shit, and C. S. Seelamantula. Training-free, single-image super-resolution using a dynamic convolutional network. *IEEE Signal Processing Letters*, 25(1):85–89, 2018.

[3] S. A. Bigdeli and M. Zwicker. Image restoration using autoencoding priors. *arXiv preprint arXiv:1703.09964*, 2017.

[4] A. Buades, B. Coll, and J.-M. Morel. A non-local algorithm for image denoising. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 60–65. IEEE, 2005.

[5] H. C. Burger, C. J. Schuler, and S. Harmeling. Image denoising: Can plain neural networks compete with bm3d? In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2392–2399. IEEE, 2012.

[6] S. G. Chang, B. Yu, and M. Vetterli. Adaptive wavelet thresholding for image denoising and compression. *IEEE transactions on image processing*, 9(9):1532–1546, 2000.

[7] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 1, 2017.

[8] C. E. Duchon. Lanczos filtering in one and two dimensions. *Journal of applied meteorology*, 18(8):1016–1022, 1979.

[9] E. Eisemann and F. Durand. Flash photography enhancement via intrinsic relighting. In *ACM transactions on graphics (TOG)*, volume 23, pages 673–678. ACM, 2004.

[10] Y. Gandelsman, A. Shocher, and M. Irani. Double-dip: Unsupervised image decomposition via coupled deep-image-priors. *arXiv preprint arXiv:1812.00467*, 2018.

[11] S. He and R. W. Lau. Saliency detection with flash and no-flash image pairs. In *European Conference on Computer Vision*, pages 110–124. Springer, 2014.

[12] D. Kim, D. Cho, D. Yoo, and I. S. Kweon. Learning image representations by completing damaged jigsaw puzzles. *WACV*, 2018.

[13] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[14] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[15] S. Lefkimmiatis. Non-local color image denoising with convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[16] Z. Li, Z. Murez, D. Kriegman, R. Ramamoorthi, and M. Chandraker. Learning to see through turbulent water. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 512–520. IEEE, 2018.

[17] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(Jan):19–60, 2010.

[18] X. Mao, C. Shen, and Y.-B. Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Advances in neural information processing systems*, pages 2802–2810, 2016.

[19] R. Mechrez, I. Talmi, F. Shama, and L. Zelnik-Manor. Learning to maintain natural image statistics. *arXiv preprint arXiv:1803.04626*, 2018.

[20] R. Mechrez, I. Talmi, and L. Zelnik-Manor. The contextual loss for image transformation with non-aligned data. *European Conference on Computer Vision (ECCV)*, 2018.

[21] G. Petschnigg, R. Szeliski, M. Agrawala, M. Cohen, H. Hoppe, and K. Toyama. Digital photography with flash and no-flash image pairs. In *ACM transactions on graphics (TOG)*, volume 23, pages 664–672. ACM, 2004.

[22] J. H. Rick Chang, C.-L. Li, B. Poczos, B. V. K. Vijaya Kumar, and A. C. Sankaranarayanan. One network to solve them all – solving linear inverse problems using deep projection models. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[23] M. S. M. Sajjadi, B. Scholkopf, and M. Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[24] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[25] A. Shocher, S. Bagon, P. Isola, and M. Irani. Internal distribution matching for natural image retargeting. *arXiv preprint arXiv:1812.00231*, 2018.

[26] R. K. Srivastava, K. Greff, and J. Schmidhuber. Training very deep networks. In *Advances in neural information processing systems*, pages 2377–2385, 2015.

[27] J. Sun, S. B. Kang, Z.-B. Xu, X. Tang, and H.-Y. Shum. Flash cut: Foreground extraction with flash and no-flash image pairs. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[28] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Computer Vision, 1998. Sixth International Conference on*, pages 839–846. IEEE, 1998.

[29] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Deep image prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[30] R. G. Vidal, S. Banerjee, K. Grm, V. Struc, and W. J. Scheirer. Ugˆ 2: A video benchmark for assessing the impact of image restoration and enhancement on automatic visual recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1597–1606. IEEE, 2018.

[31] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008.

[32] C. Wang, H. Huang, X. Han, and J. Wang. Video inpainting by jointly learning temporal structure and spatial details. *arXiv preprint arXiv:1806.08482*, 2018.

[33] J. Xie, L. Xu, and E. Chen. Image denoising and inpainting with deep neural networks. In *Advances in neural information processing systems*, pages 341–349, 2012.

[34] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 3, 2017.

[35] D. Yang and J. Sun. Bm3d-net: A convolutional neural network for transform-domain collaborative filtering. *IEEE Signal Processing Letters*, 25(1):55–59, 2018.

[36] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5485–5493, 2017.

[37] R. Zeyde, M. Elad, and M. Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010.

[38] K. Zhang, W. Zuo, S. Gu, and L. Zhang. Learning deep cnn denoiser prior for image restoration. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[39] F. Zhu, G. Chen, and P.-A. Heng. From noise modeling to blind image denoising. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 420–429, 2016.